Predicting MLB Player Salary

By Brandon Hoskins

Background

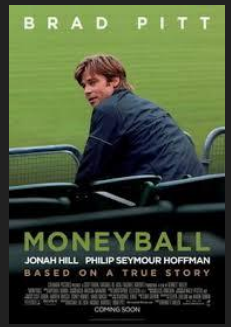
At the end of every MLB season, it seems like players are offered more and more money based on how well they performed. We have seen baseball players get paid hundreds of millions of dollars by Major League Baseball teams to play the game of baseball at the highest level. We have seen players who signed big contracts be labeled as overpaid because they did not reach the performance level expected while being paid so much money. We have also seen players outperform their contract value and rightfully deserve to be paid more money. Every MLB offseason, reporters and TV analysts try to guess how much money a free agent is going to sign for.

Background (Cont’d)

Baseball statistics are used to evaluate players based on their performance. Baseball stats have been kept and recorded since the 1870s, when professional baseball first started. Traditional stats that baseball fans use to evaluate hitters are batting average (BA or AVG), home runs (HR), runs batted in (RBIs), and hits (H). For pitchers the stats are wins (W), earned run average (ERA), strike outs (SO or K). Now, baseball statisticians and analysts have created more stats to evaluate player performance such as on base percentage (OBP), on base plus slugging (OPS), and slugging percentage (SLG) for hitters and walks plus hits per innings pitched (WHIP) and strike outs per nine (SO/9 or K/9) for pitchers. There also even more advanced stats with more factors to further evaluate players. Based on these stats, players are rewarded with bigger contracts if they perform well.

Moneyball

In the 2002 season, the Oakland Athletics GM Billy Beane had to figure out away to compete with the New York Yankees who had a much higher payroll for players than Oakland did. The A's owner was not willing to spend big time money like the Yankees, so Beane had to think of a creative way to put together a team that competes for a championship with a low player budget. Beane implemented the "Moneyball" scheme that used more advanced stats like OBP and SLG to evaluate players. Just before the 2002 season, Beane and the A's front office team searched for players that were overlooked and undervalued by other teams but had good enough statistics to be apart of a world championship baseball team. As a result of this, the 2002 A's won 103 games and set the AL record at the time of 20 wins in a row. Although, the A's did not win the World Series, they changed the game of baseball and how teams construct their roster to build championship-caliber teams.



Problem Statement

To build two regression models that can predict a player's salary based on their on-field performance and their baseball statistics for a season. Also, to build a user interactive web app with Streamlit where users can input custom baseball statistics to generate a salary prediction based on those statistics. One regression model will be used for position players only using hitting stats and the other model will be used for pitchers using pitching stats.

Data Collection

The data used in this project came from Lahman's Baseball Database on this website: http://www.seanlahman.com/baseball-archive/statistics/

I used 4 datasets from this database to help build my model:

- Batting dataset - Includes hitting stats for all players from 1871-2019.
- Pitching dataset - Includes pitching stats for all pitchers from 1871-2019.
- Salary dataset - Includes player salary for all players from 1985-2016.
- All-Star dataset - Includes data from MLB All Star Games from 1933-2019.
- Awards dataset - Includes data on award winners from 1877-2017.

Data Cleaning - Position Players Model

- Removed pitchers from the batting dataset. By merging the pitching dataset and batting dataset and removed players that showed up in both datasets.
- Dropped all rows that had stats before the 1985 season because I only had salary data from 1985 - 2016. Then, merged the salary dataset to the batting dataset.
- Dropped players who made less than the minimum MLB salary in 1985 which was $60,000. Also, dropped players that did not record an at-bat in any season.
- Added a years of experience column. How many times a player showed up in the dataset equals how many seasons that player has been in the league.
- Added an All-Star column by merging the players that were on the all star dataset to my model dataset. These players got a 1 if they were an All Star and missing values were imputed with 0 because that player was not an All Star.
- For the Awards columns, I dummified the columns in the awards dataset, then selected the awards I wanted to include for the dataset. Then, I merged the columns I wanted into my dataset. If a player won an award they got a 1 and if they did not then it was 0.

Data Cleaning - Position Players Model (cont'd)

I also wanted to add more baseball stats as features for the model:

- Batting Average (BA or AVG): the number of hits divided by at-bats
- On Base Percentage (OBP): a measure of how often a batter reaches base.
  - Formula: (Hits + Walks + Hit by Pitch) / (At Bats + Walks + Hit by Pitch + Sacrifice Flies)
- Slugging Percentage (SLG) : represents the total number of bases a player records per at bat.
  - Formula: (Hits + Doubles + (2 * Triples) + (3 * Home Runs) / At-Bats
- On Base Plus Slugging (OPS): Combines OBP and SLG to evaluate how often a player gets on base and if the player is more of a contact hitter or a power hitter.
  - Formula: OBP + SLG

I decided to use these stats because they are constantly used to evaluate and compare players.

Data Cleaning - Pitchers Model

Same process as the Position Player dataset.

The only differences are:

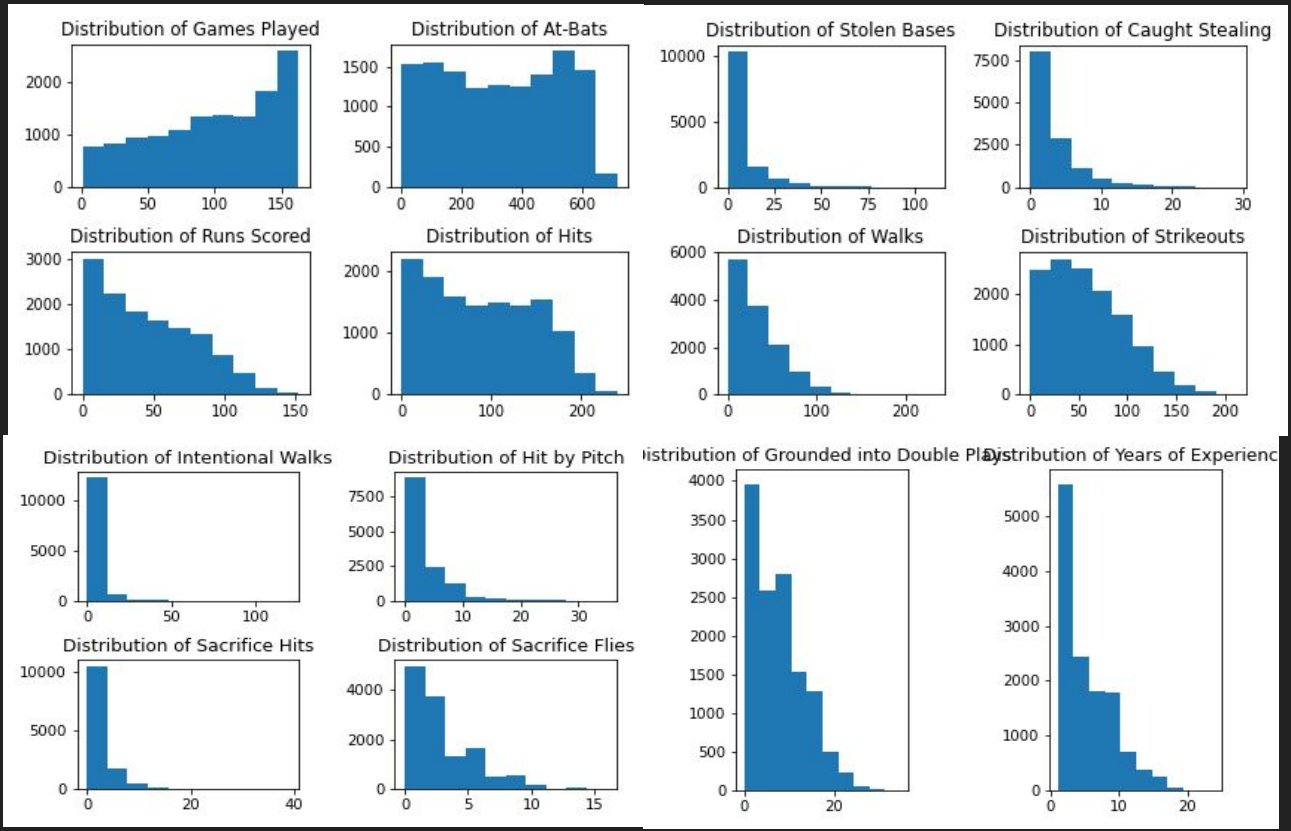
- Taking out the position players from the pitching dataset. To do this I scraped the position players pitching dataset from Baseball Reference using Beautiful Soup and merged this dataset to the pitchers dataset and took out the players that showed up on both datasets.
- Selected the awards from the awards dataset that pitchers can win.

Data Cleaning - Pitchers Model (cont'd)

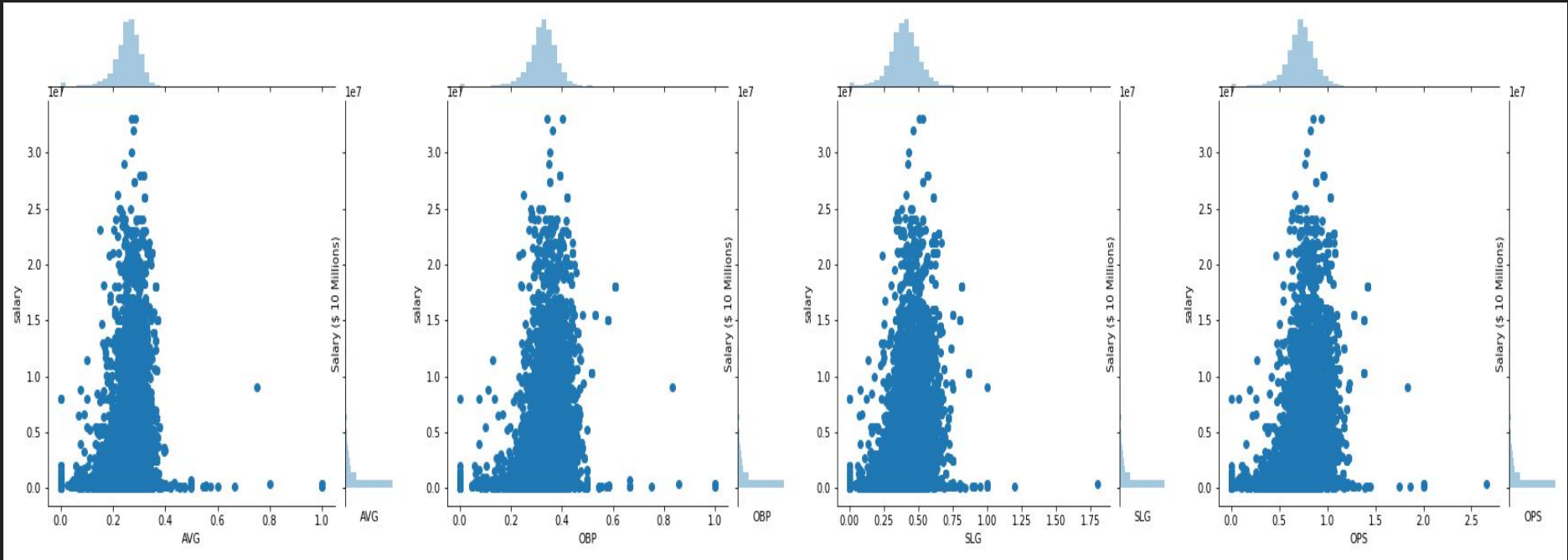
Similar to the Position Player dataset, I wanted to add in a few more pitching stats to add more features to the model.

- Walks and Hits Per Innings Pitched (WHIP): Shows how well a pitcher can keep hitters off the basepaths.
  - Formula: (Walks + Hits) / Innings Pitched
- Hits per 9 innings (H/9): The average number of hits a pitcher allows per nine innings pitched.
  - Formula: (9 * Hits) / Innings Pitched
- Home Runs per 9 innings (HR/9): The average number of home runs a pitcher allows per nine innings pitched.
  - Formula: (9 * Home Runs) / Innings Pitched
- Walks per 9 innings (BB/9): The average number of walks a pitcher allows per nine innings pitched.
  - Formula: (9 * Walks) / Innings Pitched
- Strike Outs per 9 innings (SO/9 or K/9): The average number of strike outs a pitcher gets per nine innings pitched.
  - Formula: (9 * Ks) / Innings Pitched

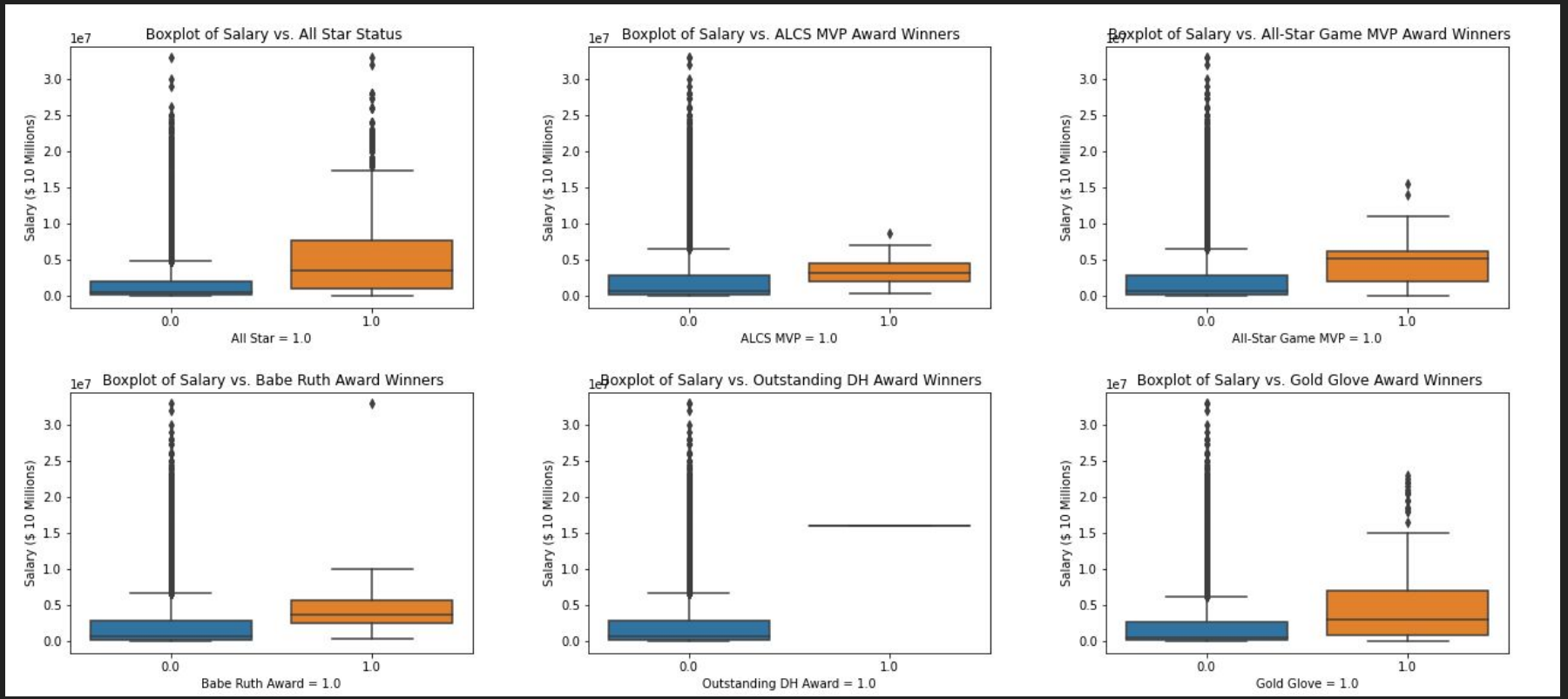
EDA- Position Players



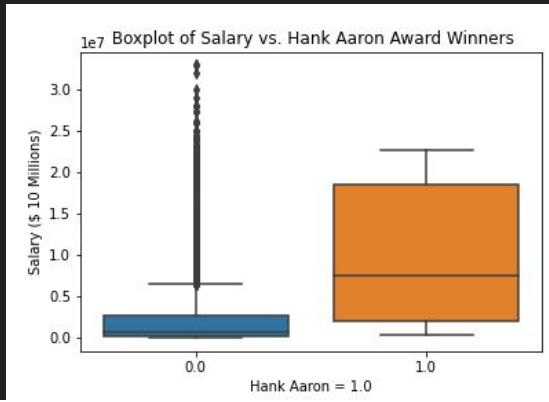
Jointplots for AVG, OBP, SLG, and OPS

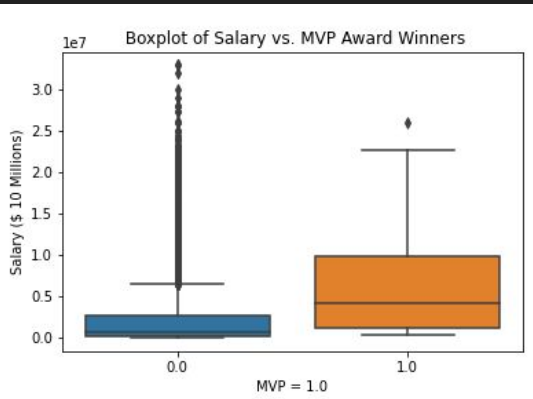


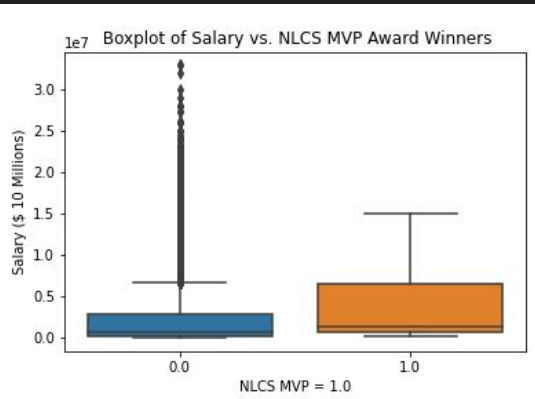
Boxplots of All-Star Status and Awards vs. Salary

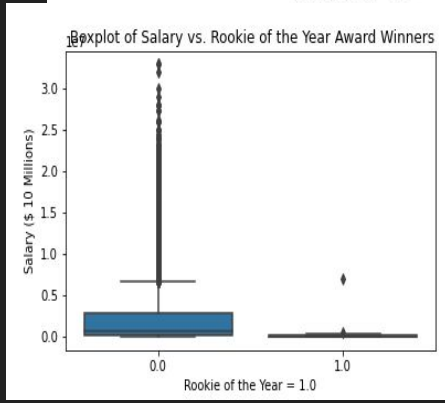


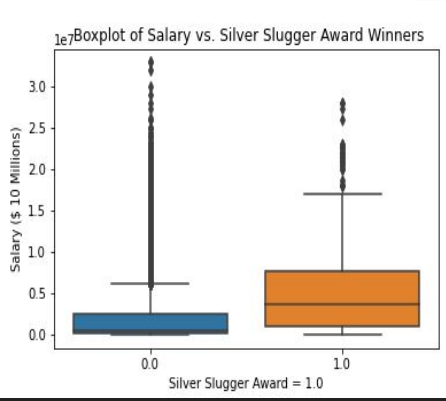
More Boxplots

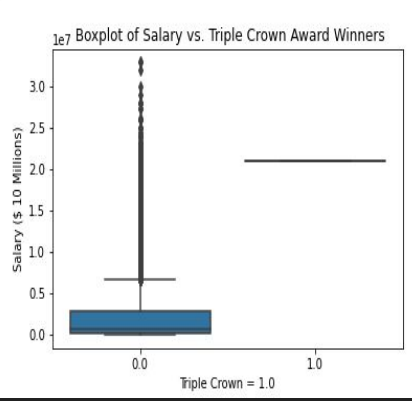


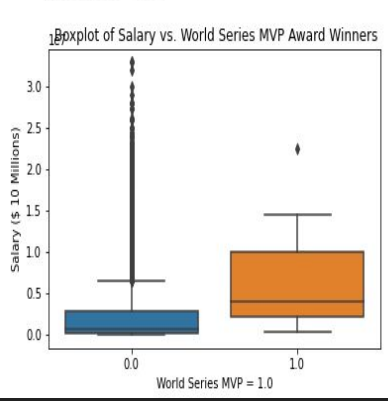




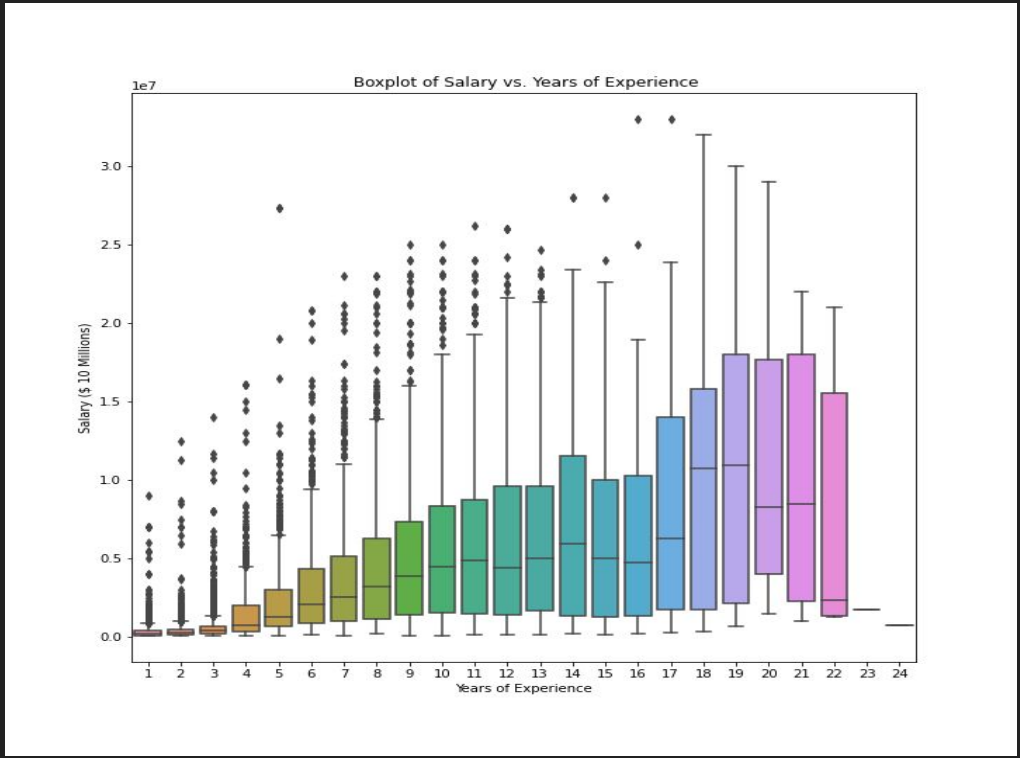




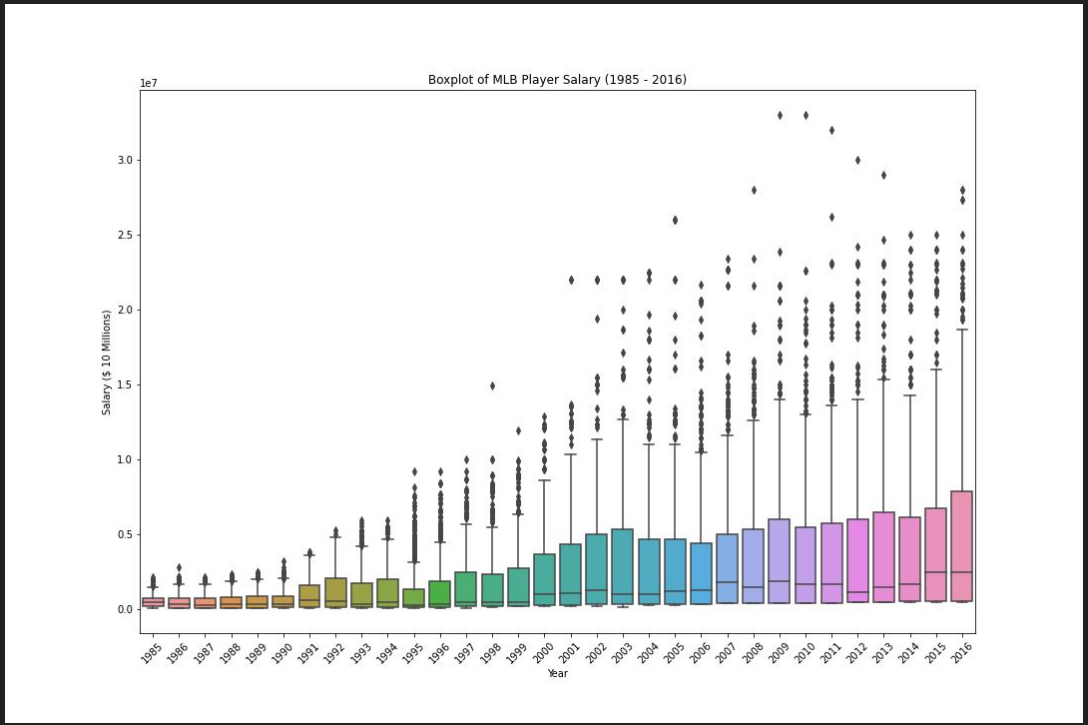




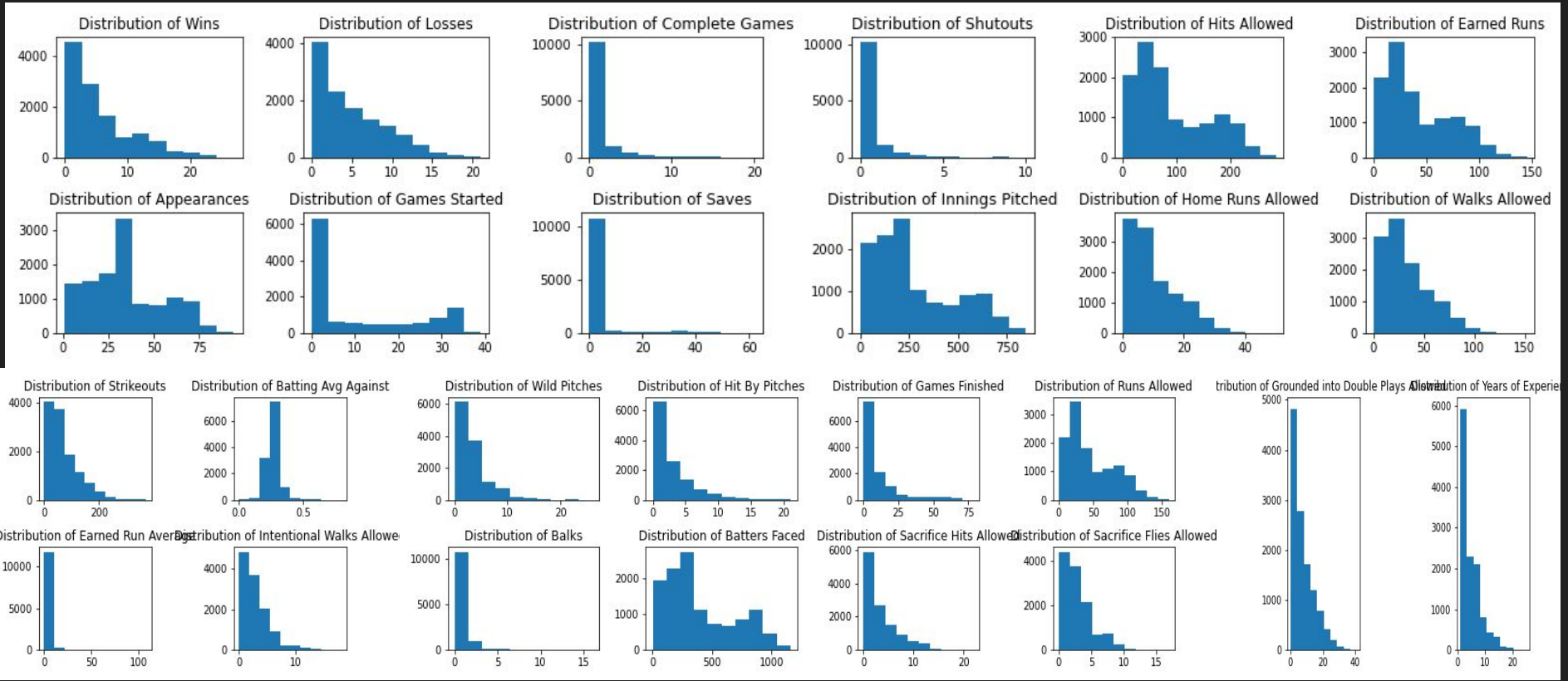
Boxplots of Player Salaries vs Years of Experience



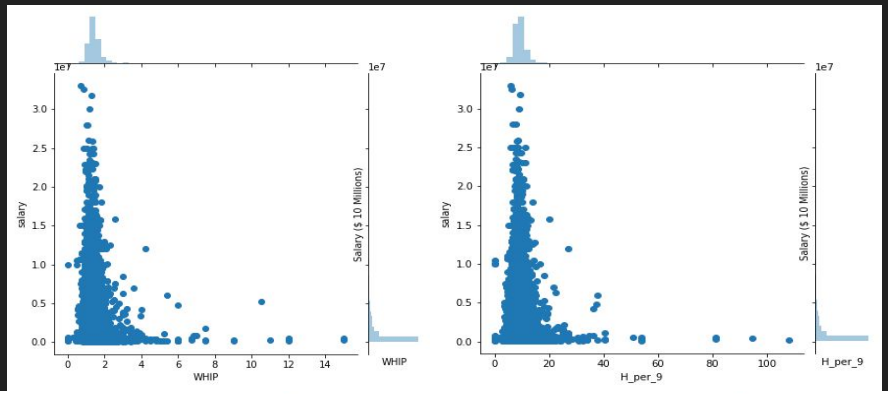
Boxplots of Player Salaries by Season

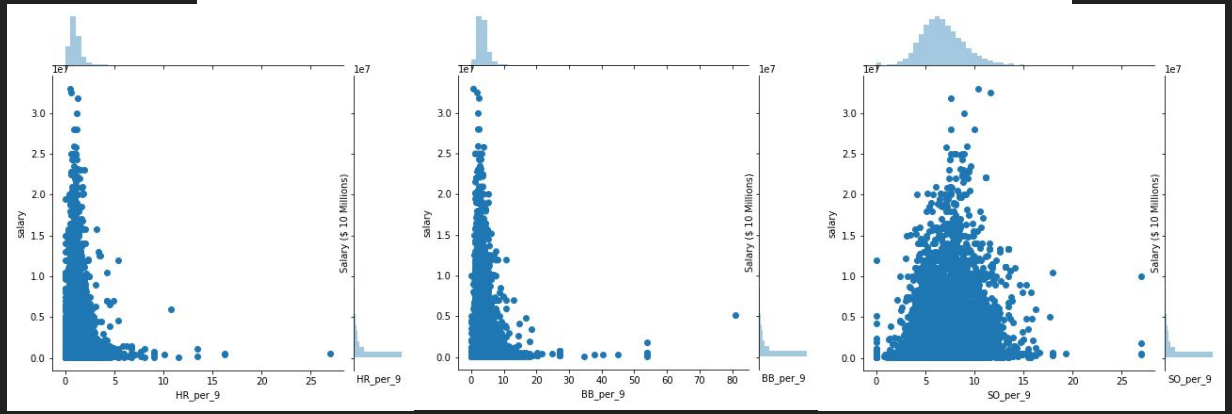


EDA - Pitchers

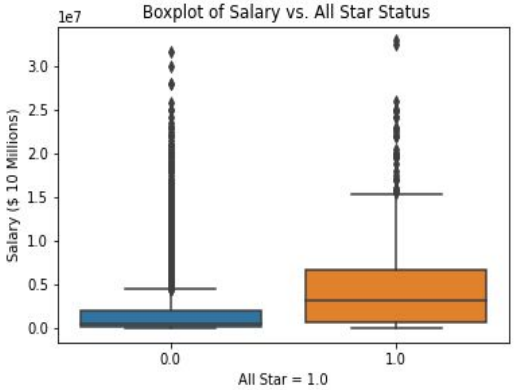


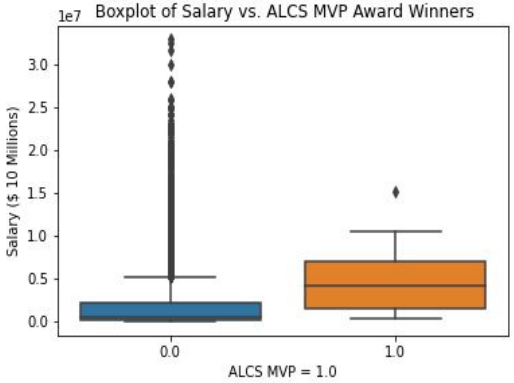
Jointplots for WHIP, H/9, HR/9, BB/9, and K/9

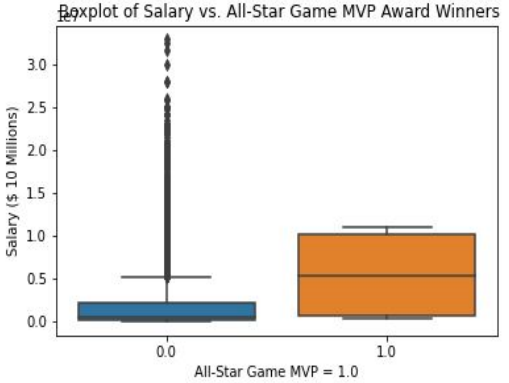


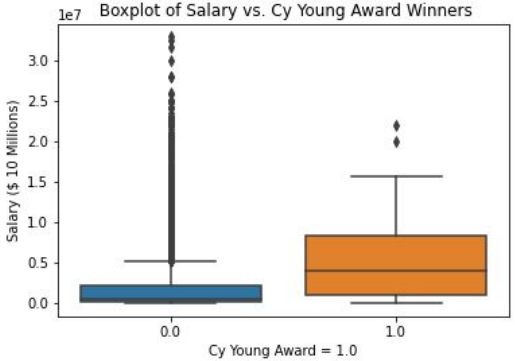


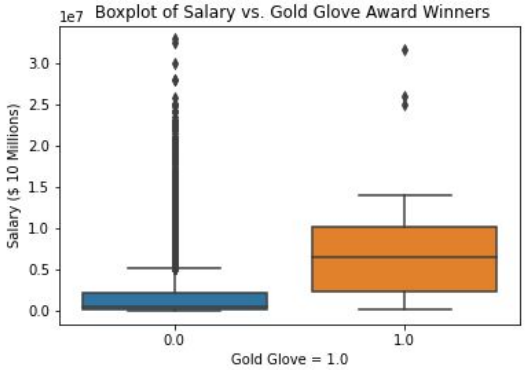
Boxplots of All-Star Status and Awards vs Salary for Pitchers

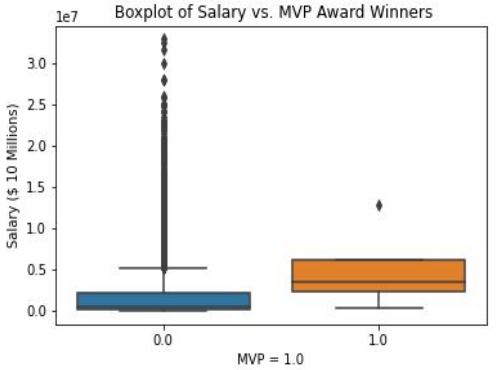




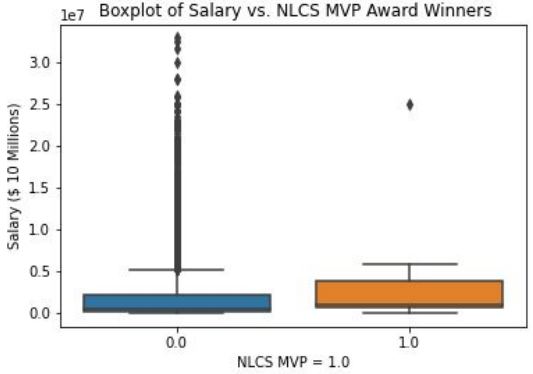


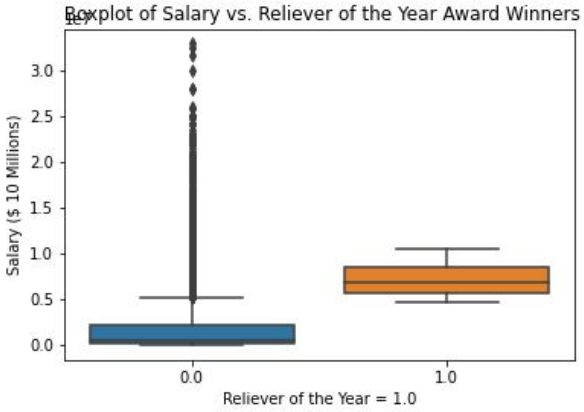


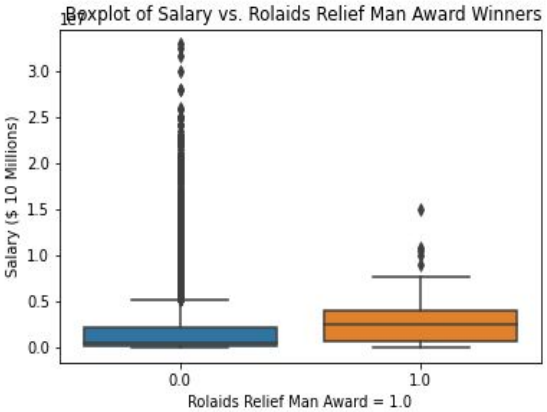


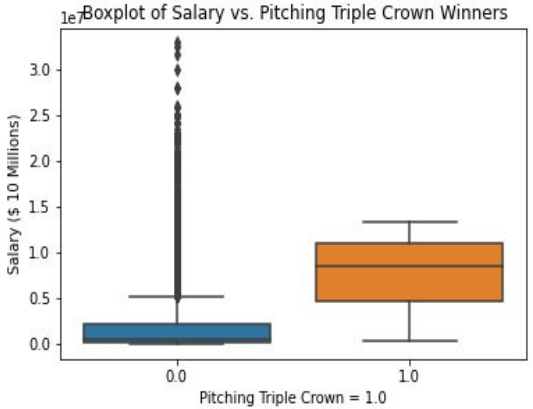


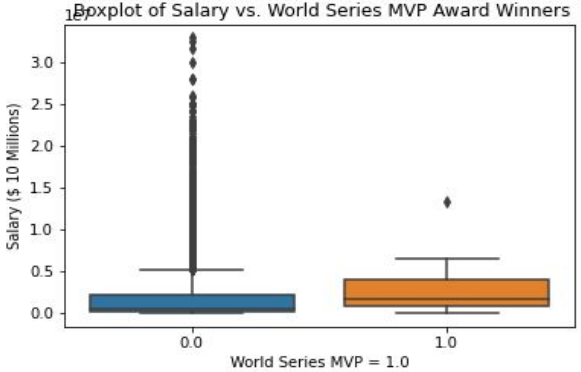
More Boxplots

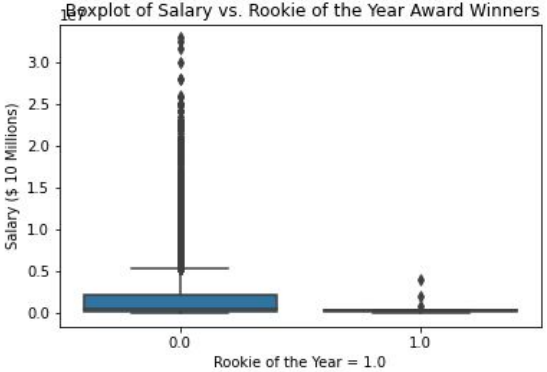




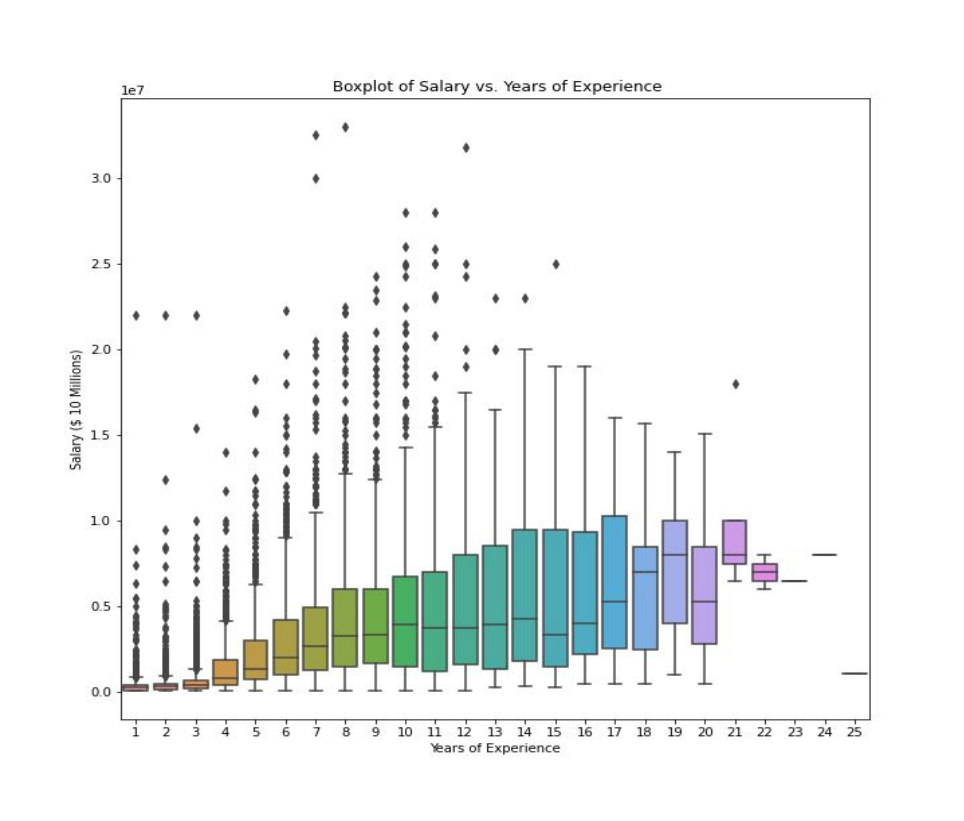




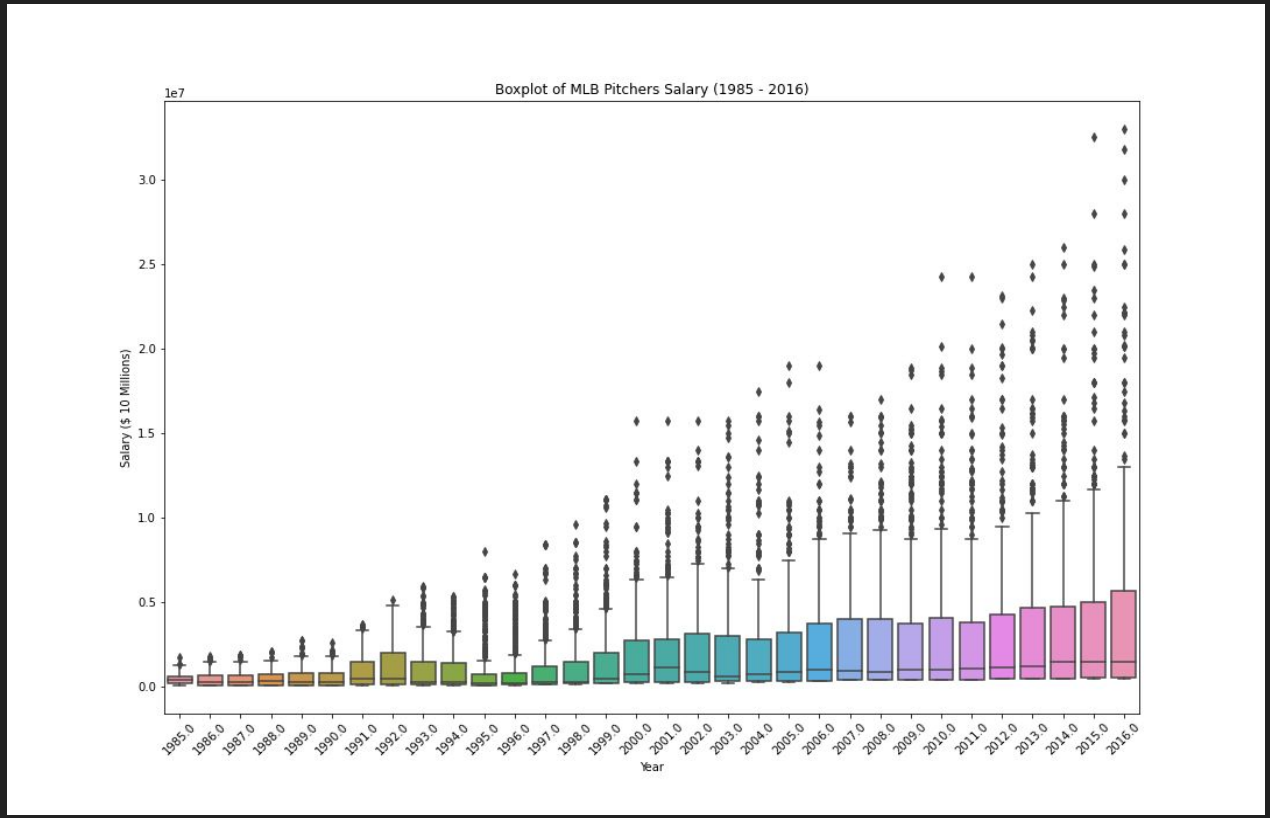




Boxplots of Pitcher Salaries vs Years of Experience

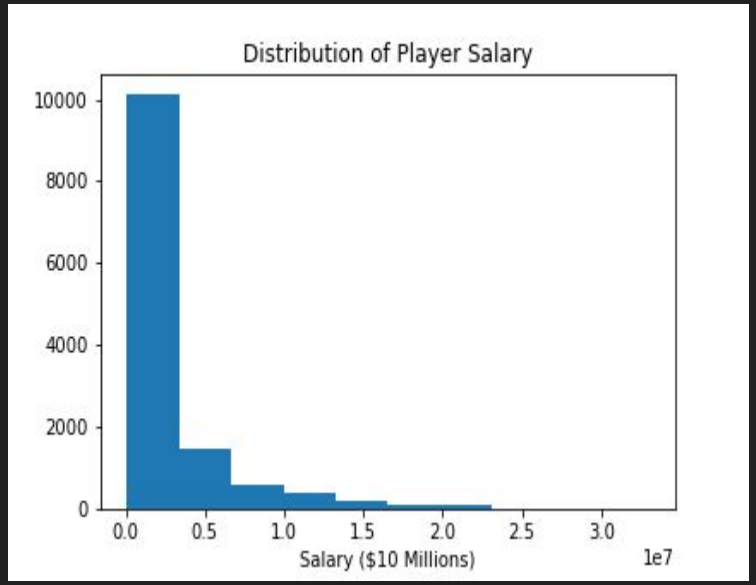


Boxplots of Pitcher Salaries Over Seasons

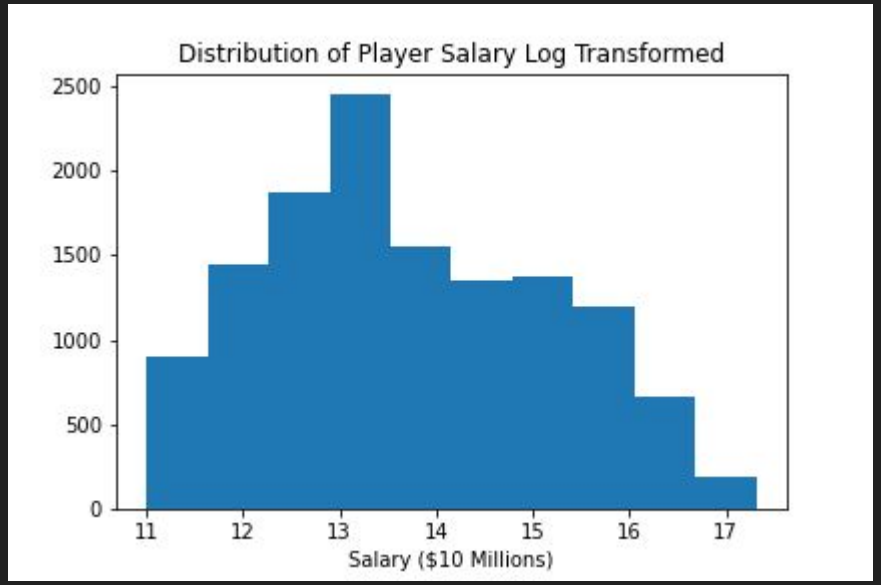


Looking at our Target Variable for Position Player Model

Our target:

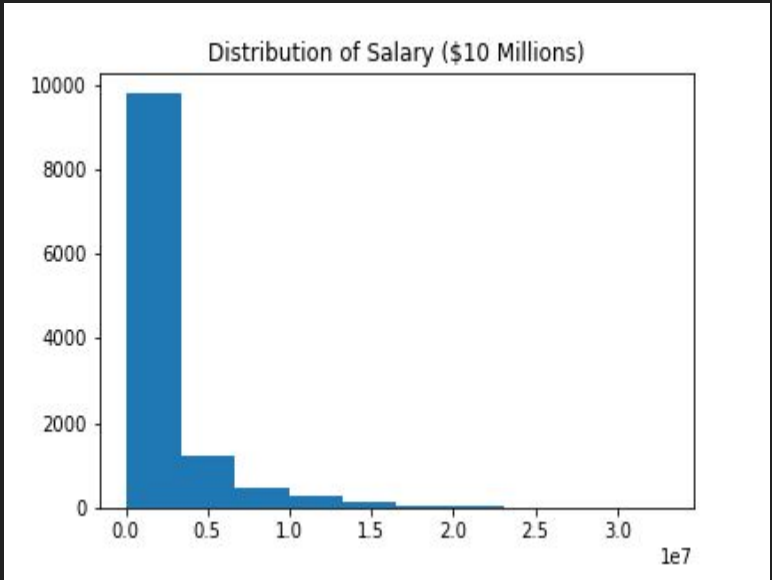


Target log transformed:

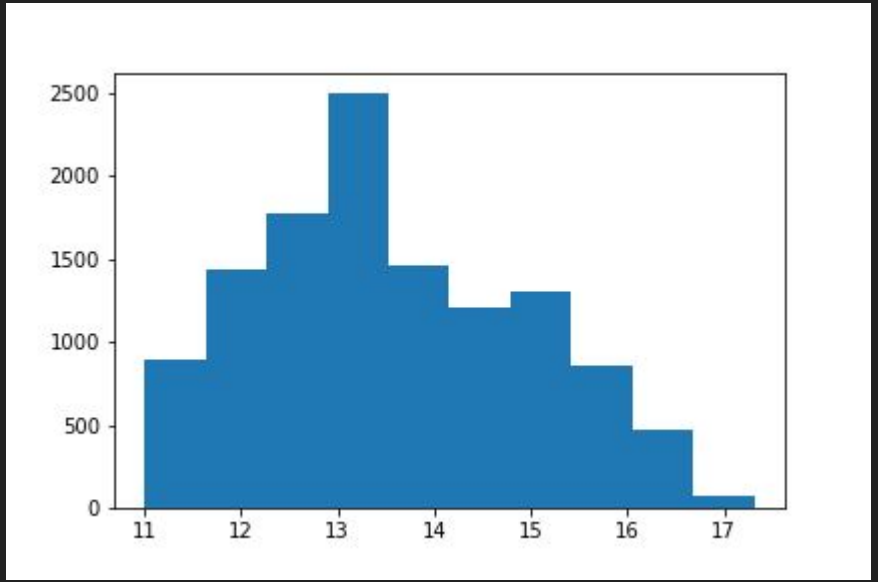


Looking at our Target Variable for Pitchers Model

Our target:



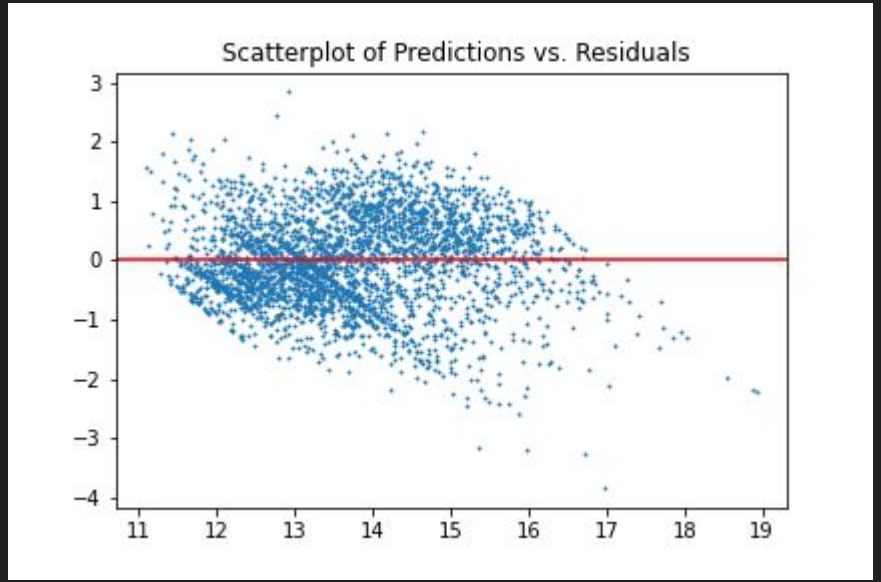
Target log transformed:



Linear Regression Model Evaluation - Position Players

- R^2 score for Training Data: 0.71
- R^2 score for Testing Data: 0.72

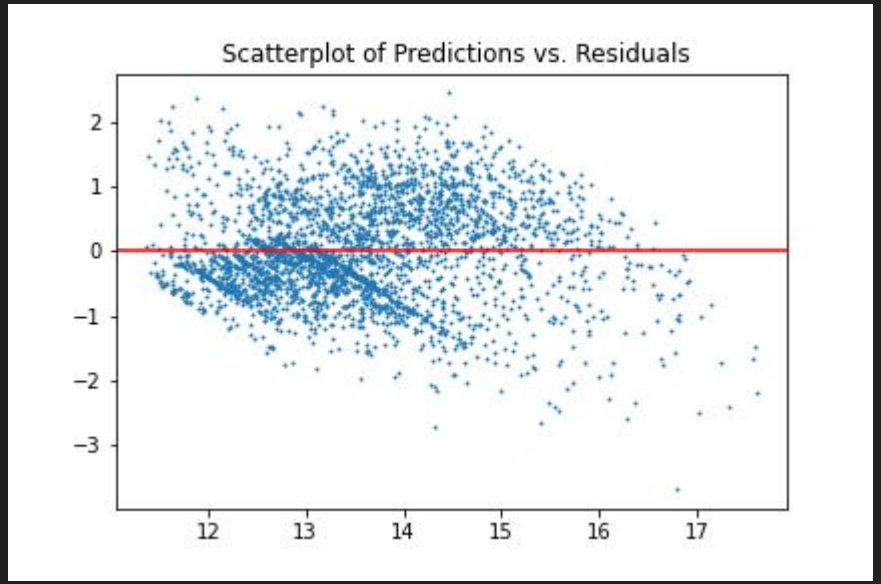
This model does an okay job of predicting player salary. However, this is our best interpretable model and it will be used for Streamlit.



Linear Regression Model Evaluation - Pitchers

- R^2 score for Training Data: 0.68
- R^2 score for Testing Data: 0.67

Similar to the position player model this model does an okay job of predicting player salary. However, this is our best interpretable model and is going to be used for Streamlit.



Streamlit App Challenge - Jacob DeGrom

Let’s look at Jacob DeGrom’s 2019 season where he was an All-Star and won his 2nd Cy Young Award.

His actual salary in 2019 was $9,000,000

Let’s see what our model predicts...

Gradient Boost Model Evaluations

Used GridSearch to find the best parameters to generate the best R^2 scores for both models.

Position Player Model:

My best parameters:

- Max_depth = 4
- Learning rate = 0.12
- N_estimators = 150

R^2 score on Training Data: 0.83

R^2 score on Testing Data: 0.84

Pitcher Model:

My best parameters:

- Max_depth = 4
- Learning rate = 0.08
- N_estimators = 125

R^2 score on Training Data: 0.81

R^2 score on Testing Data: 0.80

Conclusions and Recommendations

Although our model came close in the Streamlit challenge, there some limitations with this data that I would like to consider in the future.

- The dataset used only had data from the 1985 season to the 2016 season. I would like to have more recent data and maybe data prior to 1985. This might help the models generate better predictions.
- For features I would like to add an injury feature because the models did not factor that players could miss time because of injuries and cause them to not have the stats that we expect.
- I would like to upgrade the awards features for players that have won the award multiple times.
- I could also add even more advanced stats like WAR (Wins Above Replacement) to help the models have better scores.
- Another feature I would like to add is player positions such as shortstop and left field and fielding stats.

Sources

http://www.seanlahman.com/baseball-archive/statistics/

https://www.baseball-reference.com/friv/fieldPitch.shtml